# Face Recognition by Sparse Discriminant Analysis via Joint $L_{2,1}$-norm Minimization

Xiaoshuang Shi, Yujiu Yang, Zhenhua Guo, Zhihui Lai*

[a] *Shenzhen Key Laboratory of Broadband Network & Multimedia*
*Graduate School at Shenzhen, Tsinghua University*
*Shenzhen, 518055, China*
[b] *Graduate School at Shenzhen, Tsinghua University*
*Shenzhen, 518055, China*
[c] *Shenzhen Key Laboratory of Broadband Network & Multimedia*
*Graduate School at Shenzhen, Tsinghua University*
*Shenzhen, 518055, China*
[d] *Bio-Computing Research Center*
*Graduate School at Shenzhen, Harbin Institute of Technology University*
*Shenzhen, 518055, China*

**Abstract**

Recently, joint feature selection and subspace learning, which can perform feature selection and subspace learning simultaneously, is proposed and has encouraging ability on face recognition. In the literature, a framework of utilizing $L_{2,1}$-norm penalty term has also been presented, but some important algorithms cannot be covered, such as Fisher Linear Discriminant Analysis and Sparse Discriminant Analysis. Therefore, in this paper, we add $L_{2,1}$-norm penalty term on FLDA and propose a feasible solution by transforming its nonlinear model into linear regression type. In addition, we modify the optimization model of SDA by replacing elastic net with $L_{2,1}$-norm penalty term and present its optimization method. Experiments on three standard face

*Corresponding author
*Email address:* lai_zhi_hui@163.com (Zhihui Lai )

databases illustrate FLDA and SDA via $L_{2,1}$-norm penalty term can significantly improve their recognition performance, and obtain inspiring results with low computation cost and for low-dimension feature.
*Keywords:* $L_{2,1}$-norm, Fisher Linear Discriminant Analysis, Sparse Discriminant Analysis

## 1. Introduction

Linear Discriminant Analysis (LDA) is wildly applied to solve the supervised classification problems due to its simplicity and effectiveness. However, high-dimensional data, in which the number of predictor variables $p$ is much larger than the number of observations $n$ ($n << p$), are a trouble for LDA's applications.

To address this problem in LDA, many methods have been proposed. Generally speaking, these methods can be classified into two categories: feature selection and subspace learning. Feature selection is to select a subset of discriminative features from feature set [1], such as Linear Discriminant Feature Selection (LDFS) [2]; subspace learning, such as Penalized Discriminant Analysis (PDA) [5] and Discriminant Analysis by Gaussian mixtures (DAGM) [6, 7], is also named feature transform which transforms the original features into a learned low-dimensional features subspace [3, 4]. For the subspace learning, there is a disadvantage that the learned low-dimensional features are the combination of all original features. It is difficult to interpret which features play an important role in discriminant analysis. Thus, sparse subspace learning was proposed by using lasso constraint [8, 9] to enhance the interpretability. The representative ones are Sparse Discriminant

Analysis (SDA) [10], which was based on PDA and DAGM by adding lasso constraint, and Sparse Approximation to the Eigensubspace for Discrimination (SAED) [11] and Sparse Tensor Discriminant Analysis (STDA) [31], which used elastic net [12] to learn the sparse discriminant analysis.

Although sparse subspace learning methods can obtain encouraging ability to explore the significant features, the selected features are independent and different from each dimension. In order to discard the irrelevant features and transform the relevant ones, in the past, one intuitive way was to perform feature selection before subspace learning, but the two sub-process conducted individually would be likely to make the whole process sub-optimal [13]. Therefore, joint feature selection and subspace learning method was proposed by using $L_{2,1}$-norm penalty term [16], which can perform feature selection and subspace learning simultaneously. $L_{2,1}$-norm penalty term had been applied on graph embedding [17, 18] and a framework with encouraging discriminant ability on face recognition was proposed [13]. However, the framework is based on Brand's work [27] which cannot cover many important LDA algorithms [17], such as Fisher Linear Discriminant Analysis (FLDA) [28] and SDA. But FLDA and SDA are two representative algorithms of LDA, they are popular in many applications, especially, FLDA is a classical supervised learning for feature extraction and classification.

Motivated by above mentioned issues, in this paper, we add $L_{2,1}$-norm penalty term on FLDA and propose a feasible solution for the modified optimization type by transforming the nonlinear optimization problem into linear optimization problem. In addition, we modify the regression model of SDA by replacing elastic net with the $L_{2,1}$-norm penalty term to encourage row-

sparsity of the projective matrix. Experiments on benchmark face image data sets illustrate the effectiveness and efficiency of our approaches.

The rest of paper is structured as follows. In Section 2, we briefly review the models of LDA via joint $L_{2,1}$-norm regularization regression, FLDA and SDA. Section 3 introduces the algorithms of FLDA and SDA via $L_{2,1}$-norm optimization; Experimental results on benchmark face recognition data sets are reported and analyzed in Section 4. Finally, we give the conclusion and discuss the future study work in Section 5.

## 2. A brief review of several methods of LDA

### *2.1. Notations and definitions*

For an $n \times p$ matrix $M = (m_{ij})$, its $i$-th row and $j$-th column are denoted by $m^i$ and $m_j$ respectively. The $L_p$-norm of vector $v \in R^n$ is defined as $\|v\|_p = (\sum_{i=1}^n |v_i|^p)^{\frac{1}{p}}$ $(p \in Z^+)$. The $L_2$-norm of the matrix is defined as $\|M\|_2^2 = \sum_{i=1}^n \|m^i\|_2^2$ and the $L_1$-norm is defined as $\|M\|_1 = \sum_{i=1}^n \|m^i\|_1$, the $L_{2,1}$-norm of $M$ is defined as $\|M\|_{2,1} = \sum_{i=1}^n \|m^i\|_2$.

### *2.2. Joint $L_{2,1}$-norm regular regression in LDA*

Least square regression [19] is widely applied in linear discriminant analysis. Given training data $X = \{x_1, x_2, \cdots, x_n\} \in R^{p\times n}$, and $Y = \{y_1, y_2, \cdots, y_n\}^T \in R^{n\times c}$ are the corresponding class labels, the least square regression aims to find the projective matrix $W \in R^{p\times c}$, which can be obtained by solving the optimization problem as follow

$$\min_W \sum_{i=1}^n \left\|W^T x_i - y_i\right\|_2^2 \tag{1}$$

By adding the penalty term $\lambda\Phi(W)$ , the optimization problem becomes

$$\min_{W} \sum_{i=1}^{n} \left\| W^T x_i - y_i \right\|_2^2 + \lambda\Phi(W) \tag{2}$$

Due to the characteristic of $L_{2,1}$-norm, which can encourage row-sparsity of projective matrix [13] and the $L_{2,1}$-norm of projective matrix is convex and easily optimized [20], we replace $\Phi(W)$ with $\|W\|_{2,1}$. Thus, the regression model (2) becomes the following problem [21]

$$\min_{W} \sum_{i=1}^{n} \left\| W^T x_i - y_i \right\|_2^2 + \lambda \|W\|_{2,1} \tag{3}$$

This model had been applied on LDA in graph embedding framework to get sparse projecting matrix in [13]. Solving Eq. (3), we can get

$$W = (XX^T + \lambda G)^{-1} XY \tag{4}$$

where $G$ is a diagonal matrix with the $i$-th diagonal element equals to

$$g_{ii} = \frac{1}{2||w^i||_2} \tag{5}$$

when $||w^i|| = 0$, we can define $g_{ii} = \frac{1}{2\sqrt{||w^i||_2^2 + \varepsilon^2}}$ $(\varepsilon \to 0)$ as in [26].

Based on the Woodbury matrix identity [22], Eq. (4) can be rewritten as

$$W = G^{-1} X (X^T G^{-1} X + \lambda I)^{-1} Y \tag{6}$$

If $\lambda = 0$ , Eq. (6) is the Situation 1 in [13], otherwise, it is the Situation 2.

### *2.3. FLDA*

FLDA minimizes the within-class distance and maximizes the between-class distance, the criterion $J(W)$ can be written as follow

$$\underset{W}{max}\ J(W) = Tr\frac{W^T S_B W}{W^T S_W W} \tag{7}$$

where $S_W$ is the within-class scatter matrix and $S_B$ is the between-class scatter matrix.

By adding penalty term the $L_{2,1}$-norm of $W$, the optimization problem (7) becomes

$$\underset{W}{min}\ \widehat{J(W)} = -Tr\ \frac{W^T S_B W}{W^T S_W W} + \lambda \sum_{i=1}^{p} \left\| w^i \right\|_2 \tag{8}$$

*2.4. SDA*

SDA is proposed mainly based on PDA and DAGM by adding $L_1$-norm penalty term. It is another representative form of LDA, which is based on different principle from FLDA and has different mathematic model. For a training data set $X = \{x_1, x_2, \cdots, x_n\}^T \in R^{n\times p}$ , where $n$ represents the number of observations and $p$ is the number of predictors, $\beta \in R^{p\times d}$ represents a low-dimensional projective matrix. Its optimization model is defined as follow

$$\begin{gathered}(\widehat{\theta}, \widehat{\beta}) = arg\ \underset{\theta,\beta}{min} \sum_{i=1}^{n} \left\| y_i\theta - x_i\beta \right\|_2^2 + \lambda_1 \left\| \Omega\beta \right\|_2^2 + \sum_{j=1}^{d} \lambda_{2,j} |\beta^j|_1 \\ s.t.\ n^{-1} \left\| Y\theta \right\|_2^2 = 1\end{gathered} \tag{9}$$

where $Y \in R^{n\times c}$ is a dummy variable matrix and $c$ represents the number of class. $\theta \in R^{c\times d}$ is a scoring matrix, $\theta_{ij}$ is the score between the class $i$ and projective vector $\beta_j$. $\Omega$ is a penalization matrix.

## 3. FLDA and SDA via joint $L_{2,1}$-norm

### *3.1. FLDA via joint $L_{2,1}$-norm*

The criterion $\widehat{J(W)}$ is constituted by $J(W)$ and an $L_{2,1}$-norm penalty term, it is a nonlinear regression type and cannot be directly solved as Eq. (3). In order to get a feasible solution, we divide its optimization process into two steps:

(1) Transform the nonlinear criterion $J(W)$ into a linear optimization model.

(2) Find the optimal solution $W$.

If $W^T S_W W = I$ , $J(W)$ can be transformed into the following optimization problem

$$\max_W J(W) = Tr\frac{W^T S_B W}{W^T S_W W} = \max_{W^T S_W W = I} Tr\ W^T S_B W \tag{10}$$

$S_W$ is a symmetric matrix, based on the singular value decomposition (SVD) [29], $S_W$=UD$U^T$, so $W^T U D U^T W = I$. If we define

$$Y = D^{\frac{1}{2}} U^T W \tag{11}$$

Then $Y^T Y = I$, and

$$W = U D^{-\frac{1}{2}} Y \tag{12}$$

Eq. (12) suggests that $S_W$ must be positive-definite matrix. However, in the condition of small sample size, $S_W$ is a singular matrix, thus, we transform it into a nonsingular matrix by PCA before SVD as fisherface [25], it can become

$$\widehat{S_W} = W_{PCA}^T S_W W_{PCA} \tag{13}$$

Therefore, Eq. (10) can be rewritten as

$$\max_{W_f} J(W_f) = \max_{W_f} Tr \frac{W_f^T W_{PCA}^T S_B W_{PCA} W_f}{W_f^T W_{PCA}^T S_W W_{PCA} W_f} = \max_{W_f^T \widehat{S_W} W_f = I} Tr W_f^T \widehat{S_B} W_f \tag{14}$$

where $\widehat{S_B} = W_{PCA}^T S_B W_{PCA}$, $W_f = W_{PCA}^T W$;

Then we apply SVD on $\widehat{S_W}$ instead of $S_W$, and then get the $Y$ and $W_f$ according to Eq. (11) and Eq. (12).

Substituting $W_f$ obtained by Eq. (12) into Eq. (14), we can get

$$\begin{gathered} \hat{Y} = arg \max_{Y} Y^T D^{-\frac{1}{2}} U^T \widehat{S_B} U D^{-\frac{1}{2}} Y \\ s.t.\ Y^T Y = I \end{gathered} \tag{15}$$

The solution of $Y$ is the eigenvector of $D^{-\frac{1}{2}} U^T \widehat{S_B} U D^{-\frac{1}{2}}$

Based on Eq. (11), if we define $X = UD^{\frac{1}{2}}$, then $X^T W_f = Y$ is a linear system problem, which may behave one of three possible ways: (1) infinite solutions; (2) a single unique solution; (3) no solution, see [13]. The most popular way to solve this problem is to apply the penalty term, by adding $L_{2,1}$-norm penalty term, the criterion $X^T W_f = Y$ can be written as the linear optimization model (3) and the solution of $W_f$ is Eq. (4). There are two situations for the solution of $W_f$ in Eq. (4):

(1) $\lambda = 0$, the solution of $W_f$ is

$$W_f = G^{-1} X (X^T G^{-1} X)^{-1} Y \tag{16}$$

In this situation, the linear system problem results in the infinite number of solutions.

Substituting $X = UD^{\frac{1}{2}}$ into Eq. (16), we can get

$$W_f = G^{-1}UD^{\frac{1}{2}}(D^{\frac{1}{2}}U^TG^{-1}UD^{\frac{1}{2}})^{-1}Y \tag{17}$$

Thus, we can get $W$ based on $W = W_{PCA} * W_f$;

In summary, we present this situation for obtaining the optimal $W$ of the criterion $\widehat{J(W)}$ in Algorithm 1.

**Algorithm 1:** FLDA via $L_{2,1}$-norm (Situation 1) (L21FLDA)

**Initialize:** $G_0$=I, $t = 0$;

Compute $U$ and $D$ based on SVD of $W_{PCA}^T S_W W_{PCA}$

Compute $Y$ based on the eigenvalue decomposition of $D^{-\frac{1}{2}}U^T\widehat{S_B}UD^{-\frac{1}{2}}$

**repeat**

Compute $W_{ft+1} = G_t^{-1}UD^{\frac{1}{2}}(D^{\frac{1}{2}}U^TG_t^{-1}UD^{\frac{1}{2}})^{-1}Y$

Compute $G_{t+1}$ based on $W_{ft+1}$

t=t+1;

**until** $W_f$ converge

**Construct the final projection:** $W = W_{PCA} * W_f$

(2) $\lambda \neq 0$, the solution is in Eq. (6).

Substituting $X = UD^{\frac{1}{2}}$ into Eq. (6), the optimal $W_f$ can be obtained as follow

$$W_f = G^{-1}UD^{\frac{1}{2}}(D^{\frac{1}{2}}U^TG^{-1}UD^{\frac{1}{2}} + \lambda I)^{-1}Y \tag{18}$$

It includes two cases that the linear system problem leads to one single solution or no solution.

Then we can obtain $W$ based on $W = W_{PCA} * W_f$;

This situation for obtaining the optimal $W$ of the criterion $\widehat{J(W)}$ is presented in Algorithm 2.

**Algorithm 2:** FLDA $L_{2,1}$-norm (Situation 2) (L21FLDA)

**Initialize:** $G_0$=I, t=0;

Compute $U$ and $D$ based on SVD of $W_{PCA}^T S_W W_{PCA}$

Compute $Y$ based on the eigenvalue decomposition of $D^{-\frac{1}{2}} U^T \widehat{S_B} U D^{-\frac{1}{2}}$

**repeat**

Compute $W_{ft+1} = G_t^{-1} U D^{\frac{1}{2}} (D^{\frac{1}{2}} U^T G_t^{-1} U D^{\frac{1}{2}} + \lambda I)^{-1} Y$

Compute $G_{t+1}$ based on $W_{ft+1}$

t=t+1;

**until** $W_f$ converge

**Construct the final projection:** $W = W_{PCA} * W_f$

### *3.2. SDA via joint $L_{2,1}$-norm*

For the optimization model of SDA, we replace its penalty term with the $L_{2,1}$-norm of $\beta$, thus its optimization model becomes

$$\begin{aligned} (\widehat{\theta}, \widehat{\beta}) = arg \min_{\theta,\beta} \textstyle\sum_{i=1}^{n} n^{-1} \left\| y_i\theta - x_i\beta \right\|_2^2 + \left\| \Omega\beta \right\|_{2,1} \\ s.t.\ n^{-1} \left\| Y\theta \right\|_2^2 = 1 \end{aligned} \tag{19}$$

where $\Omega$ is a penalty diagonal matrix.

If we define $D = n^{-1} Y^T Y$, $D$ is a symmetric positive-definite matrix, then $\theta^T D \theta = I$. Next, we define $\theta^* = D^{\frac{1}{2}} \theta$ , then $\theta^{*T} \theta^* = I$, substituting them into the regression model (19), it becomes the following regression model

$$(\widehat{\theta^{*T}}, \widehat{\beta}) = arg \min_{\theta^*,\beta} n^{-1} \sum_{i=1}^{n} \left\| y_i D^{-\frac{1}{2}} \theta^* - x_i \beta \right\|_2^2 + \|\Omega\beta\|_{2,1} \quad s.t.\ \theta^{*T}\theta^* = I \tag{20}$$

Fixed $\theta^*$, Eq. (20) can be viewed as Eq. (3) by replacing $\Omega$ with $\lambda$, therefore, the solution of $\beta$ becomes

$$\beta = (X^T X + \Omega G)^{-1} X^T Y D^{-\frac{1}{2}} \theta^* \tag{21}$$

Substituting Eq. (21) into Eq. (20), it becomes the following problem

$$\max_{\theta^*} Tr\, \theta^{*T} D^{-\frac{1}{2}} Y^T X (X^T X + \Omega G)^{-1} X^T Y D^{-\frac{1}{2}} \theta^* \quad s.t.\ \theta^{*T}\theta^* = I \tag{22}$$

Based on Theorem 3 and 4 of [9], $\theta^* = UV^T$, where $U$ and $V$ can be obtained by SVD of $D^{-\frac{1}{2}} Y^T X \beta$ . Finally, the solution of $\theta = D^{-\frac{1}{2}} UV^T$

Depending on whether the value of $\Omega$ is zero, the solution of Eq. (19) also can be divided into two situations, but we put them together as Algorithm 3 for simplicity.

## 4. Experiments

In order to evaluate the performance of Algorithm 1 and Algorithm 2 (L21FLDA) and Algorithm 3 (L21SDA), we applied them on three standard face databases and compared them with three algorithms, such as fisherface [25], SDA [10] and SSLDA [24]. In addition, we presented the recognition results of the algorithms FSSL (LDA) in [13]. In this paper, for better distinguishing the algorithms with $L_{2,1}$-norm penalty term, we name FSSL (LDA) as L21LDA. Moreover, according to the difference between Eq. (17) and Eq. (18), we

**Algorithm 3:** SDA via $L_{2,1}$-norm (L21SDA)

**Initialize:** $G_0$=I, $t = 0$, $\theta_0 = D^{-\frac{1}{2}} I_{1:c,1:d}$;

**repeat**

Fixed $\theta_t$, compute $\beta_{t+1} = (X^T X + \Omega G_t)^{-1} X^T Y \theta_t$

Fixed $\beta_{t+1}$, compute $U$ and $V$ based on the SVD of $D^{-\frac{1}{2}} Y^T X \beta_{t+1}$

Compute $\theta_{t+1} = D^{-\frac{1}{2}} U V^T$

Compute $G_{t+1}$ based on $\beta^{t+1}$

$t = t + 1$

**until** $\beta$ converge

also made a comparison between situation (1) $\lambda = 0$, $\Omega = 0$, and situation (2) $\lambda \neq 0$, $\Omega \neq 0$.

### *4.1. Data sets*

In our experiments, we used the following three standard face databases:

**ORL face database** contains 400 face images of 40 human subjects under a dark homogenous ground with the subjects in an upright, frontal position. In this experiment, all images are chosen and each face image is resized to 32×32 pixels, which means each face image can be presented by a 1024-dimensional vector, and the images of one human subject are presented in the top of Fig 1.

**Extended Yale-B face database** contains 16128 face images of 38 human subjects under 9 poses and 64 illumination conditions. We choose the frontal pose and use all the images under different illumination in this experiment, so there are 2414 face images in total. All the face images are manually aligned and cropped, and they are also resized to 32×32 pixels. Ten face

Fig 1: Face images of three databases, these images from top to bottom respectively belong to: ORL database, Extended YaleB database, PIE database

images of one human subject are shown in the middle of Fig 1.

**CMU PIE face database** contains 41368 face images of 68 human subjects under 13 different poses and 43 illumination conditions, and with 4 different expressions. In this experiment, 11554 images are selected in all face images, and they are manually cropped and resized to 32×32 pixels. Ten images of one human are displayed in the bottom of Fig 1.

### *4.2. Parameter settings*

The sets of training images were randomly selected in each database, and the remained images were used for testing. On ORL database, $p = [3, 5, 7]$, $p = [10, 20, 30]$ in Extended YableB database and PIE database, $p$ is the number of training images of each person. We repeated this process 50 times and calculated the mean accuracy and computation time. Generally, each image would be described by a low-dimensional vector before recognition. In our experiments, the vector of each image was reduced to $c - 1$ in all algorithms in order to better compare their running time. $c$ was the number of face classes.

For L21LDA, before performing LDA, the dimensionality was firstly re-

duced to $n - c$ by PCA as in [25]. In L21FLDA, the dimensionality was determined by the rank of $S_W$, we chose the eigenvectors corresponding to the eigenvalues greater than $10^{-4}$ to construct $\widehat{S_W}$ in our experiments. For L21LDA, L21FLDA and L21SDA, they all could be divided into two situations: (1) $\lambda = 0$ and $\Omega = 0$, we ran this case on ORL database; (2) $\lambda \neq 0$ and $\Omega \neq 0$, for L21LDA and L21FLDA, we used cross validation by searching the grid $\{0.001, 0.005, \cdots, 1\}$ to select the best $\lambda$; for L21SDA, we searched the grid $\{1, 2, \cdots, 10\}$; for SSLD, we tuned the parameter by searching the grid $\{10, 20, \cdots, 100\}$ according to [24]; for SDA, we searched the grid $\{-10, -20, \cdots, -100\}$. Besides, for ORL database, in which $p$ was small, we adopted leave-one cross validation; for Extended YaleB and PIE databases, 5-fold cross validation was adopted.

### *4.3. Results*

The results of the experiments are shown in Tables 1-4. Table 1 shows the recognition accuracy of situation (1) on ORL database. Form Table 1 and Table 2, we can see that the situation (2) $\lambda \neq 0$ and $\Omega \neq 0$ can obtain better recognition accuracy than situation (1). Tables 2-4 present the recognition rate of situation (2) on three different databases.

As shown in Tables 2-4, first, L21FLDA, L21SDA and L21LDA respectively have better recognition accuracy than fisherface, SDA and SSLDA in most cases, which demonstrates the effectiveness and efficiency of $L_{2,1}$-norm penalty term. The reason for this is that using $L_{2,1}$-norm penalty term can make them perform feature selection and subspace learning simultaneously, which can improve subspace learning and encourage row-sparsity [13]. As a side note, the biggest difference between L21LDA and SSLDA is that they

Table 1: Face recognition accuracy on ORL database ( $\lambda = 0$, $\Omega = 0$)

| Data set | 3 training | | 5 training | | 7 training | |
|---|---|---|---|---|---|---|
| | Acc (%) | Time | Acc (%) | Time | Acc (%) | Time |
| L21LDA | 79.31±2.74 | 0.890 | 91.35±2.02 | 1.171 | 94.63±2.23 | 1.613 |
| **L21SDA** | 61.11±0.1 | 0.852 | 77.56±2.12 | 1.168 | 84.15±2.83 | 1.555 |
| **L21FLDA** | 81.99±2.59 | 0.631 | 93.24±1.78 | 0.461 | 96.47±1.65 | 0.306 |

have different penalty terms. Second, among three algorithms with $L_{2,1}$-norm penalty term, the computation cost of L21FLDA is the smallest in most cases. The main reasons are that L21FLDA is unary linear regression type and the $L_{2,1}$-norm of projective matrix is convex and easily optimized, and it can quickly converge to equilibrium point [20], and the main reason why L21FLDA has less computation time than L21LDA is that the dimension after using PCA was smaller in L21FLDA than in L21LDA. In order to better understand the reason of less computation cost for L21FLDA, we present the iteration process of three algorithms with $L_{2,1}$-norm penalty term in Fig. 2. Furthermore, performing PCA sometimes would severely increase the computation cost, that is why L21SDA (without PCA) has less computation cost than L21FLDA and L21LDA as $p = [10, 20]$ in Table 4.

### *4.4. Recognition accuracy vs. dimension*

In this subsection, we present the correlation between the recognition accuracy and the dimension. Fig. 3 shows the performance of six algorithms on ORL, Extended YaleB and PIE databases with 5 training, 20 training and 20 training samples respectively. We ran the algorithms for 20 times independently and then computed the average accuracy. The x-axis is di-

Table 2: Face recognition accuracy on ORL database ( $\lambda \neq 0$, $\Omega \neq 0$)

| Data set | 3 training | | 5 training | | 7 training | |
|---|---|---|---|---|---|---|
| | Acc (%) | Time | Acc (%) | Time | Acc (%) | Time |
| fisherface | 83.32±1.97 | 0.078 | 92.86±1.98 | 0.065 | 95.25±2.39 | 0.105 |
| SDA | 80.24±3.85 | 28.42 | 87.81±4.99 | 29.61 | 91.18±3.58 | 31.88 |
| SSLDA | 83.07±1.66 | 1.950 | 92.32±1.93 | 3.937 | 94.87±2.41 | 5.491 |
| L21LDA | 80.74±2.81 | 0.886 | 92.44±1.158 | 1.158 | 95.70±1.77 | 1.639 |
| **L21SDA** | 83.60±2.70 | 0.924 | 92.40±1.74 | 1.401 | 94.87±1.87 | 2.097 |
| **L21FLDA** | 82.08±2.59 | 0.632 | 93.29±1.72 | 0.466 | 96.57±1.69 | 0.322 |

Table 3: Face recognition accuracy on Extended YaleB database

| Data set | 10 training | | 20 training | | 30 training | |
|---|---|---|---|---|---|---|
| | Acc (%) | Time | Acc (%) | Time | Acc (%) | Time |
| fisherface | 87.21±1.15 | 0.270 | 91.24±0.85 | 0.676 | 86.96±1.26 | 1.580 |
| SDA | 68.77±8.68 | 17.44 | 73.42±9.45 | 41.48 | 76.46±8.83 | 72.72 |
| SSLDA | 84.69±1.29 | 4.437 | 92.24±1.08 | 9.671 | 94.93±0.73 | 15.34 |
| L21LDA | 88.50±1.13 | 2.323 | 95.58±0.78 | 6.215 | 98.04±0.53 | 13.92 |
| **L21SDA** | 85.86±1.23 | 5.256 | 94.23±0.74 | 16.06 | 97.03±0.56 | 25.70 |
| **L21FLDA** | 83.75±1.49 | 1.210 | 94.40±0.91 | 2.215 | 97.32±0.85 | 4.382 |

Table 4: Face recognition accuracy on PIE database

| Data set | 10 training | | 20 training | | 30 training | |
|---|---|---|---|---|---|---|
| | Acc (%) | Time | Acc (%) | Time | Acc (%) | Time |
| fisherface | 78.40±0.92 | 1.718 | 84.66±0.55 | 2.138 | 91.98±0.36 | 2.058 |
| SDA | 72.85±2.67 | 83.97 | 78.40±4.16 | 200.8 | 79.58±6.56 | 249.4 |
| SSLDA | 84.76±0.90 | 36.95 | 91.64±0.52 | 51.53 | 93.66 ±0.33 | 82.33 |
| L21LDA | 85.33±0.76 | 30.53 | 92.21±0.36 | 41.66 | 94.05±0.29 | 55.54 |
| **L21SDA** | 78.54±0.83 | 18.80 | 87.56±1.23 | 26.68 | 91.72±1.61 | 37.39 |
| **L21FLDA** | 86.38±0.63 | 28.19 | 92.18±0.24 | 27.72 | 94.56±0.24 | 26.50 |

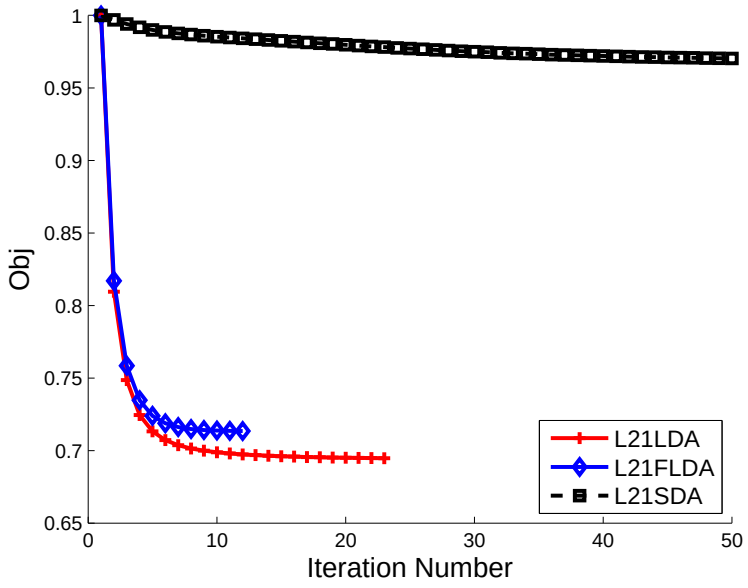


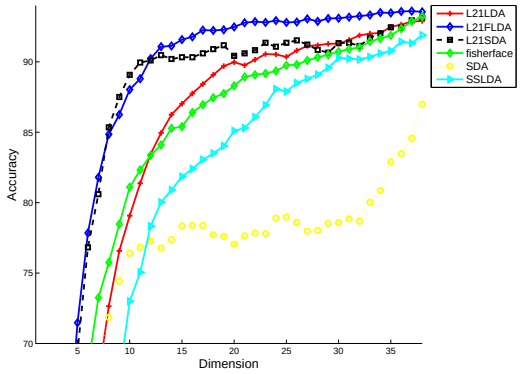

Fig 2: Iteration process of (3), (8) and (9) with $p = 20$ on Extended YaleB database. X−label is the iteration number, Y−label represents the value of object function which has been normalized in order to make comparison; similar iteration process can be observed for other $p$ and databases.

mensionality and y-axis represents the recognition accuracy. It displays the detailed changes of recognition accuracy vs. the dimension variations.

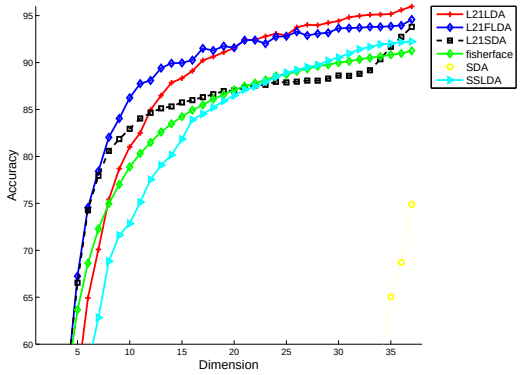
Based on Fig. 3, we can see that L21FLDA and L21SDA have superiority on recognition performance in low-dimensional subspace with other algorithms. It suggests that L21FLDA and L21SDA can be applied on the classification problems of low-dimension samples. For L21FLDA, the main reason is that using $L_{2,1}$-norm penalty term improves the performance of fisherface which has well performance in low-dimensional subspace [25]. For L21SDA, one main reason is that SVD makes the score weights of focus on low-dimensional subspace, which can be inferred from Eq. (22), it is the similar trend as for the weights of according to Eq. (21), this reason is also applicable to SDA which can quickly get temporary stabilization (see Fig. 3 (a) and (c)); the other main reason is that $L_{2,1}$-norm can encourage row-sparsity and rank the importance of the features [30], thus the selected features are nearly the same in each dimension. Similar phenomenon can be observed when we use a different number of training samples, due to the space limit, we do not show them.

## 5. Conclusion

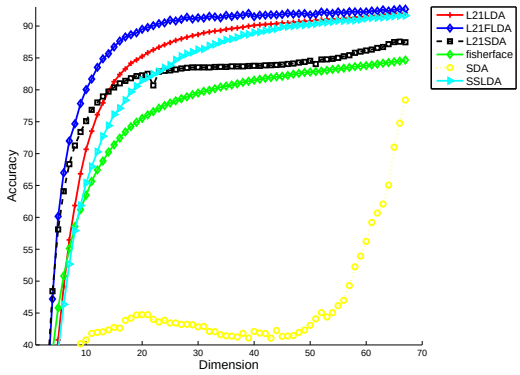
In this paper, we modify the optimization model of FLDA by adding $L_{2,1}$-norm penalty term and present a feasible solution by transforming its nonlinear optimization regression type into a linear optimization problem. Meanwhile, we propose a new optimization type for SDA by using $L_{2,1}$-norm penalty term and present its optimization process. Experiments on benchmark databases demonstrate the effectiveness and efficiency of our algorithms. We




(a) ORL




(b) Extended YaleB




(c) PIE

Fig 3: Face recognition accuracy with dimension reduction change. (a)ORL database (b) Extended YaleB database (c) PIE database

can get comparable results with fisherface and SDA, using $L_{2,1}$-norm penalty term significantly improve their recognition performance. In addition, L21FLDA has the least computation cost among three algorithms with $L_{2,1}$-norm penalty term. Furthermore, the proposed methods can get much better results in low-dimensional subspace. In the future work, we will study other representative algorithms such as Marginal Fisher Analysis (MFA) which cannot be covered in joint feature selection and subspace learning framework either.

**Acknowledge**

This work was supported by the Natural Science Foundation of China (NSFC) (No. 61101150), and the National High-Tech Research and Development Plan of China (863) (No. 2012AA09A408). Shenzhen special fund for the strategic development of emerging industries (Grant No. JCYJ201208311657 30901), the Natural Science Foundation of China (Grant Nos. 61203376, 61005005, 61071179, 61125305, 61375012), the General Research Fund of Research Grants Council of Hong Kong (Project No.531708), the China Postdoctoral Science Foundation under Project 2012M510958 and 2013T60370, the Guangdong Natural Science Foundation under Project S2012040007289, and Shenzhen Municipal Science and Technology Innovation Council (Nos. JC201005260122A, JCYJ20120613153352732 and JCYJ201206131348 43060, JCYJ20130329152024199).

## References

[1] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, The Journal of Machine Learning Research. 3 (2003) 1157-1182.

[2] M. Masaeli, J.G. Dy, G.M. Fung, From transformation-based dimensionality reduction to feature selection, in: Proceedings of the International Conference on Machine Learning, 2010, pp. 751-758.

[3] X. Niyogi, Locality preserving projections, in: Neural information processing systems, 2004, p. 153.

[4] X. He, D. Cai, S. Yan, H.J. Zhang, Neighborhood preserving embedding, in: IEEE International Conference on Computer Vision, 2005, pp. 1208-1213.

[5] T. Hastie, A. Buja, R. Tibshirani, Penalized discriminant analysis, The Annals of Statistics. (1995) 73-102.

[6] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, Journal of the Royal Statistical Society. Series B (Methodological). (1996) 155-176.

[7] T. Hastie, R. Tibshirani, B. Andreas, Flexible discriminant and mixture models, Statistics and neural networks: advances at the interface. (1999) 1-23.

[8] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological). (1996) 267-288.

[9] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, Journal of computational and graphical statistics. 15 (2) (2006) 265-286.

[10] L. Clemmensen, T. Hastie, D. Witten, B. Ersb?ll, Sparse discriminant analysis, Technometrics. 53 (4) (2011) 406-413.

[11] Z.H. Lai, W.K. Wong, Z. Jin, J. Yang, Y. Xu, Sparse approximation to the eigensubspace for discrimination. 23 (22) (2012) 1948 - 1960.

[12] H. Zou, T. Hastie, Regression shrinkage and selection via the elastic net, with applications to microarrays, Journal of the Royal Statistical Society: Series B. 67 (2003) 301-320.

[13] Q. Gu, Z. Li, J. Han, Joint feature selection and subspace learning, in: Proceedings of the Twenty-Second international joint conference on Artificial, 2011, pp. 1294-1299.

[14] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology). 68 (1) (2006) 49-67.

[15] A. Argyriou, T. Evgeniou, M. Pontil, Convex multi-task feature learning, Machine Learning. 73 (3) (2008) 243-272.

[16] G. Obozinski, B. Taskar, M.I. Jordan, Joint covariate selection and joint subspace selection for multiple classification problems, Statistics and Computing. 20 (2) (2010) 231-252.

[17] S. Yan, D. Xu, B. Zhang, H.J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, Pattern Analysis and Machine Intelligence, IEEE Transactions on. 29 (1) (2007) 40-51.

[18] C. Hou, F. Nie, D. Yi, Y. Wu, Feature selection via joint embedding learning and sparse regression, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2011, pp. 1324-1329.

[19] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, The Annals of Statistics. 32 (2) (2004) 407-499.

[20] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint $L_{2,1}$-norms minimization, in: Proceedings of the Neural Information Processing Systems, 2010, pp. 1813-1821.

[21] R. Jenatton, J.Y. Audibert, F. Bach, Structured variable selection with sparsity-inducing norms, The Journal of Machine Learning Research. 12 (2011) 2777-2824.

[22] G.H. Golub, C.F. Van Loan, Matrix computations, JHU Press, 2012.

[23] S.P. Boyd, L. Vandenberghe, Convex optimization, Cambridge university press, 2004.

[24] D. Cai, X. He, J. Han, Spectral regression: A unified approach for sparse subspace learning, in: IEEE International Conference on Data Mining, 2007, pp. 73-82.

[25] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, Pattern Analysis and Machine Intelligence, IEEE Transactions on. 19 (7) (1997) 711-720.

[26] F. Nie, H. Wang, H. Huang, C. Ding, Early active learning via robust representation and structured sparsity, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2013, pp. 1572-1578.

[27] M. Brand, Continuous nonlinear dimensionality reduction by kernel

eigenmaps, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2003, pp. 547-554.

[28] M. Welling, Fisher linear discriminant analysis, Department of Computer Science, University of Toronto, 2005.

[29] O. Alter, P.O. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, Proceedings of the National Academy of Sciences. 97 (18) (2000) 10101-10106.

[30] C.P. Hou, F.P. Nie, D.Y. Yuan, Y. W, Feature selection via joint embedding learning and sparse regression, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2011, pp. 1324-1329.

[31] Z.H. Lai, Y. Xu, J. Yang, J.H. Tang, D. Zhang, Sparse Tensor Discriminant Analysis, IEEE Transactions on Image Processing. 22 (10) (2013) 3904-3915.